REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
31-May-2016	Final Technical Report	01-09-2011 to 31-March-2016

4. TITLE AND SUBTITLE

Synthetic Teammates as Team Players: Coordination of Human and Synthetic Teammates

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-11-1-8444

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Nancy J. Cooke, Mustafa Demir, Nathan McNeese

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Cognitive Engineering Research Institute
7565 E. Eaglecrest Dr. #100
Mesa, AZ 85207

8. PERFORMING ORGANIZATION REPORT NUMBER

RE2016844_01

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Marc Steinberg, Paul Bello
Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release, distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This project is part of a larger effort that focuses on human-automation coordination in the context of the development, integration, and validation of a computational cognitive model that acts as a full-fledged synthetic teammate on an otherwise all-human team. The team performs a small command-and-control task (i.e., team control of an Unmanned Aerial System; UAS). The research integrated the synthetic teammate model into the CERTT II (Cognitive Engineering Research on Team Tasks II) testbed in order to empirically address these research questions: 1) What is the nature of coordination and collaboration (within human or mixed human-synthetic teams) in command and control (C2) settings; 2) How well does the synthetic teammate function as part of a human-synthetic agent team; and 3) What do deficiencies in synthetic teammate interactions with human teammates reveal about human-automation coordination needs?

15. SUBJECT TERMS

synthetic teammate, human-autonomy teaming, team coordination

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Nancy Cooke, Ph.D.
U	U	U	UU	39	19b. TELEPHONE NUMBER *(include area code)* (480) 988-7306

Grant or Contract # N000141110844

Synthetic Teammates as Team Players: Coordination of Human and Synthetic Teammates

Final Technical Report for Period: September 1, 2011-March 31, 2016

PI: Nancy J. Cooke
480-988-1000
ncooke@cerici.org

Authors: Nancy J. Cooke, Mustafa Demir, Nathan McNeese

Cognitive Engineering Research Institute, Mesa, AZ

Date Prepared: May 31, 2016

TABLE OF CONTENTS

1.0 OVERVIEW OF PROJECT

This project is part of a larger effort that focuses on human-automation coordination in the context of the development, integration, and validation of a computational cognitive model that acts as a full-fledged synthetic teammate on an otherwise all-human team. The modeling effort was conducted by the AFRL (Air Force Research Lab) and a CRADA (Cooperative Research and Development Agreement) was established to support AFRL's collaboration with CERI (Cognitive Engineering Research Institute). The three-agent team performs a small command-and-control task (i.e., team control of an Unmanned Aerial System; UAS).

The research described in this report integrates the synthetic teammate model into the CERTT-II (Cognitive Engineering Research on Team Tasks II) testbed in order to test the validity of the synthetic teammate in terms of how well it functions as part of a human-synthetic agent team (vs. an object of supervisory control) and to empirically address these research questions: What is the nature of coordination and collaboration (within human or mixed human-synthetic teams) in command and control settings, and what do deficiencies in synthetic teammate interactions with human teammates reveal about human-automation coordination needs.

2.0 OBJECTIVES

The objectives of this effort are:

- To better understand the cognitive requirements for synthetic agents to interface with human team members
- To increase understanding of team coordination of human and autonomous components
- To achieve various applied objectives:
  - To reduce resources (i.e., participants) needed for experiments and training events
  - To enable experimental control of team dynamics
  - To facilitate team training
  - To serve as team members or coaches in mixed human-agent teams

3.0 BACKGROUND

In this section we provide a description of the synthetic task environment that provides the context for humans to team with the synthetic teammate. We discuss prior work in that context and the theory of interactive team cognition that resulted from findings of empirical studies in that context and others. Finally we describe the computational cognitive model (developed by AFRL) that is the synthetic teammate.

3.1 The Context: Unmanned Aerial System – Synthetic Task Environment (UAS-STE)

The Cognitive Engineering Research on Team Tasks Unmanned Aerial System-Synthetic Task Environment (CERTT-UAS-STE; Figure 1) was designed to be both a flexible research platform and a realistic task environment (Cooke, Rivera, Shope, & Caukwell, 1999; Cooke & Shope, 2004, 2005). In the recent synthetic teammate experiment, an updated version of this platform, CERTT-

II, was used with a view to research, development, and evaluation of the synthetic teammate, as well as to simulate teamwork aspects of UAS operations. The features of this updated version of the platform include 1) the ability for human and synthetic teammates to communicate via text chat, and 2) eight hardware consoles: four consoles for four team members and four consoles for experimenters in order to oversee the simulation, inject perturbations, and make observations (Cooke & Shope, 2004, 2005). More specifically, CERTT II features are listed:

- Capability for Distributed Interactive Simulation (DIS)
- Ability to vary team sizes of 2, 3, or 4 participants as well as teams of teams
- Capability of a team to control multiple UAV's
- Software includes training modules with tests
- Experimenter access to participant screens
- Experimenters can take control of participant applications
- Ability to disable select communication channels or insert noise in voice communication channels
- Authoring: Easy to change start-up parameters and waypoint library to define a scenario
- Possible to insert team situation awareness roadblocks into scenario (e.g. audio, video, ad-hoc targets, etc.)
- Text chat interface
- Extensive measurement capabilities
- Demonstration mode
- META/VR world imagery

The CERTT-II task environment consists of multiple 40-minute missions wherein reconnaissance photographs of certain target waypoints must be obtained by three heterogeneous teammates: 1) Air Vehicle Operator (AVO or pilot) – controls the UAS's heading, altitude, and airspeed; 2) Data Exploitation, Mission Planning, and Communications (DEMPC or navigator) – generates a dynamic flight plan and issues speed and altitude restrictions; and 3) Payload Operator (PLO or sensor operator) – monitors sensor equipment, negotiates with the AVO on speed and altitude in order to take a good photo of the target waypoints. Communication within the three-agent UAS teams occurred over a text-based communications system. In order to take a good photo for each target waypoint, the coordination among the three team members needed to follow an optimal coordination sequence: *Information-Negotiation-Feedback (INF)*: the navigator (DEMPC) provides the *information* about the upcoming target waypoint to the pilot (AVO); the AVO *negotiates* with the photographer (PLO) about an appropriate altitude and airspeed for the target waypoints and required camera settings; and finally, the PLO sends *feedback* to the AVO and the DEMPC about whether they have a good photo or not for the current target waypoint (Cooke, Gorman, Duran, & Taylor, 2007). Due to the time-sensitive nature of the task, adherence to the INF coordination sequence is vital in order to maintain stable communication within the group and avoid any communication failures that would adversely affect team performance. Note that there are many possible temporal patterns of these coordination elements that maintain the same sequencing.

3.2 Prior related work

Since 1997 when the CERTT Lab was first developed with DURIP (AFOSR) funding and updated in 2010 with DURIP (ONR) funds, there have been 10 major studies of team performance in the

operation of UAS, and more recently, team performance when interacting with autonomy - not autonomous vehicles, but autonomous teammates. These experiments have led to a number of discoveries including:

1) Empirical results demonstrating the importance of team interaction over shared knowledge (e.g., Gorman & Cooke, 2011)
2) The theory of Interactive Team Cognition (Cooke, Gorman, Myers, & Duran, 2013)
3) Metrics of team cognition that focus on interaction including the quantification of dynamics of team coordination and communication (e.g., Gorman, Amazeen, & Cooke, 2010 and summarized in the next section)
4) Perturbation methods for training adaptive teams (e.g., Gorman, Cooke, & Amazeen, 2010)

Figure 1a.

Figure 1b.

Figure 1a: Experimenter consoles and **1b:** Participant consoles of the CERTT II environment.

3.3 Interactive Team Cognition

Empirical results in the CERTT Lab have contributed to a theory of Interactive Team Cognition (Cooke, Gorman, Myers, & Duran, 2013) that emphasizes the importance of teammate interaction over individuals' knowledge, skills, and abilities. ITC takes team cognition beyond team knowledge by postulating team interactions as cognitive processes that are more directly tied to team effectiveness than knowledge. If team members properly distribute their knowledge within and among the team, yet lack effective coordination (due to minimal or failed interaction), then it is likely the team will fail to meet their objective. Thus, team interaction (i.e., team communication and coordination) is team cognition and, as such, dynamical communication and coordination patterns can be indicators to monitor team cognition (Cooke, Gorman, Myers, & Duran, 2013).

Therefore, team cognition can be assessed by focusing on team interaction (i.e., communication and coordination) among the team members (Cooke, Gorman, Duran, et al., 2007). Patterns of communication flow and content can serve as indices of team coordination (Cooke & Gorman, 2009) and team situation awareness (Gorman et al., 2005; Gorman, Cooke, & Winner, 2006). There are several studies (Gorman et al., 2006; Cooke, Gorman, Pedersen, et al., 2007; Cooke,

Gorman, & Kiekel, 2008; Cooke & Gorman, 2009; Gorman, Amazeen, & Cooke, 2010) which consider interactions among team members to be an important predictor of team performance.

The strengths and limits of the synthetic teammate provided a critical test of this theory and at the same time pushed the bounds of the ACT-R cognitive modeling architecture which has been focused on small-scale models of individual cognition.

CERTT research has also led to the development of a number of innovative measures of team effectiveness. Composite performance measures based on individual and team outcome can be used to track skill acquisition and retention at the team and individual levels. Communication among team members also provides input for measures with particular attention paid to patterns of communication flow (Cooke & Gorman, 2009). Coordination in this task has been defined as the timely and adaptive exchange of information among teammates and in the CERTT context is measured in terms of patterns of timing of specific information exchange events. Dynamical systems parameters have been applied to these patterns as they unfold over time to indicate coordination stability, flexibility, and resilience (Gorman, Amazeen, et al., 2010). Finally, team situation awareness has been measured in terms of the timely and adaptive perception of change in the environment and corrective action on the part of only those team members who are absolutely necessary (Gorman, Cooke, & Winner, 2006). These measures have been leveraged in this effort, as we compared all-human teams in the CERTT-II testbed to a team with two humans and a synthetic teammate.

Finally, there are a host of empirical findings from previous research that inform our current experimental designs and research questions. Using the measures described above, we have observed the development of team coordination over time, its adaptation to novel events, and its resilience to environmental perturbations. Specifically, for this effort in the context of a three-person team or a human–synthetic agent team, we examine the cognitive and social mechanisms that underlie this coordination behavior and hypothesize (pertinent to our scientific objectives):

- Differences in team coordination will coincide with changes in communication, leadership, team situation awareness, interruption, and information push and pull behaviors. These patterns of correlated behaviors will suggest mechanisms of team coordination.
- The synthetic teammate will be more successful independently than as a team player.
- There will be human-synthetic agent team deficits and the pattern of such deficits will further implicate mechanisms of team coordination. Specifically, we are hypothesizing that the limitations will include: inappropriate timing of information push and pull, poor back-up behavior for teammates, lack of adaptation to teammate differences, and inappropriate timing of interruptions.

3.4 The Synthetic Teammate

The synthetic teammate was developed using the Atomic Components of Thought–Rational (ACT-R) computational cognitive architecture (Anderson, 2007). The synthetic teammate acts as the AVO or pilot in the UAV-STE. ACT-R has been under continuous development for several decades and is now capable of accurately reproducing human microcognitive processes (e.g., memory retrieval, skill acquisition, etc.) Without detailing ACT-R, cognition revolves around the

interaction between a central procedural system and several peripheral modules. There are modules for vision, motor capabilities, memory, one for storing the model's intentions for completing the task (i.e., the control state), and a module for storing the mental representation of the task at hand (problem state, see Figure 2). (For more detail on ACT-R, see Anderson, 2007).

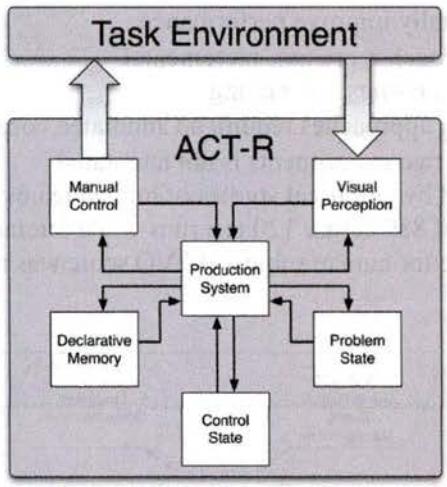


Figure 2. The modules of the ACT-R 6.0 computational cognitive architecture. Adapted from Anderson (2007)

AFRL chose ACT-R for two important reasons. First, there was an abundance of ACT-R expertise on AFRL's PALM team. Second, and more importantly, ACT-R provided a good foundation to investigate how macrocognitive processes (e.g., meta-cognition, situation assessment/awareness, etc.) affect microcognitive processes, and vice versa. Because ACT-R provides good quantitative predictions of human performance across many microcognitive processes, using them as a foundation for developing macrocognitive processes helped to uncover how micro and macro processes interact within complex task environments involving human-automation coordination.

Synthetic teammate development was managed through a *divide-and-conquer* strategy across a set of components, combined with a *synthesis* strategy for component integration. To support synthesis and cognitive plausibility of the four major components, they were all developed within the ACT-R architecture. The major components include: 1) language comprehension, 2) language generation and dialog management, 3) task behavior, and 4) the situation component. The situation component provides context to each of the other components to influence their behavior, and is updated in return as a result of their behavior (see Figure 3).

The synthetic teammate is unique in several respects. First, it is a functional system (unlike typical cognitive models) that adheres to well-established cognitive constraints (unlike typical AI systems). In addition the synthetic teammate model:

1. Is one of the largest cognitive models ever built
   - 2400 productions
   - 58,000 word mental lexicon
   - Near size of human mental lexicon

2. Achieved 95% accuracy in part of speech tagging un-tokenized text (random sample)
   - State-of-the-art is 97% accuracy with tokenized text
3. Handles ill-formed and misspelled input
4. Outperformed a state-of-the-art system on random sample from CERTT text chat corpus that was collected in two previous studies
5. Has the ability to incrementally improve performance
   - Machine learning approaches are non-incremental
6. Doesn't require an annotated corpus for training
   - Most machine learning approaches require an annotated corpus
   - Text chat corpus from two experiments is not annotated
7. Its development was guided by empirical studies of human teams
8. Achieved an overall score of 889 across 120 test runs while interacting with lightweight agents. This score exceeds the average for human subjects (AVO score was mean of 767 in the experiment reported here).

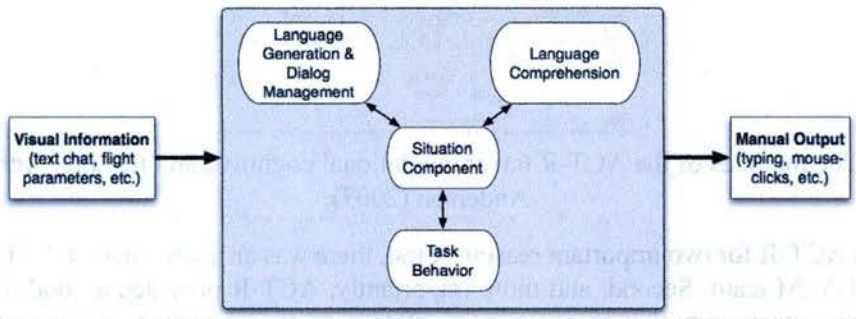


Figure 3. Functional components of the synthetic teammate

For more information on the synthetic teammate model see (Ball et al., (2010), and Rodgers, Myers, Ball, & Freiman (2013)). Validation of the synthetic teammate model consisted of integrating the synthetic teammate in the CERTT II testbed and including it as a team member in an experiment with two human teammates. Data collected in this context on team and individual performance, team process, and team and individual cognition was compared to a condition in which three humans interact to achieve the same goals and another three-human team in which the AVO was high-functioning.

4.0 ACCOMPLISHMENTS

In this project we leveraged previous work in the CERTT Lab, the theory of Interactive Team Cognition, and the synthetic teammate to examine human-synthetic teammate teaming. Synthetic teammate development was an ongoing process that paralleled work in the lab. Thus there were early experiments conducted with all-human teams to collect baseline data with which to evaluate the synthetic teammate. The integration of the synthetic teammate into the task environment was also a nontrivial part of this effort. In the following sections we describe two all-human baseline studies, the integration of the synthetic teammate into the lab and the final validation study.

4.1 Experiment 1: Voice vs. Chat Communications

The first baseline experiment was conducted to establish a baseline for a new text chat mode of communication. A decision was made to migrate communications in the CERTT Lab from intercom-based voice communications to text-based communications. This was done for two reasons: 1) The synthetic teammate could handle text better than voice, and 2) Command and control environments are increasingly using text-based communications.

Because text chat is not a transient signal like voice and because communications can occur asynchronously, there is a possibility that coordination among teammates using text chat could be altered. Specifically the coordination score should be impacted by the asynchronous nature of communication. If coordination is made more difficult, performance is also likely to be negatively impacted in this task. Not only did this experiment address questions about coordination and text chat, but it also provided a baseline against which to compare future performance and coordination data when the synthetic teammate was part of a team.

Also, given the preponderance of text-based communications in our society and its adoption in time critical military and civilian contexts, the comparison of text versus voice as modes of communication is relevant and of increasing importance. By many accounts (Weeks, Kelly, & Chapanis (1974) and Baltes, Dickson, Sherman, Bauer, & LaGanke (2002), the use of text chat may not be the best mode of communication in time-pressured circumstances. The purpose of the experiment was to investigate how text-based communications affect team performance and coordination within the UAS-STE. Based on previous research, we hypothesized that teams communicating with text would coordinate differently from teams communicating using voice and that teams communicating with voice would perform the task better than those using text.

4.1.1 Method

4.1.1.1 Participants. Twenty, three person teams comprised of college students and the general population of the Mesa, Arizona area voluntarily participated in the experiment. Individuals were compensated for their participation by payment of $10.00 per hour with each of the three team-members on the highest performing team receiving a $100.00 bonus. The majority of the participants were males, representing 76% of the sample. Individuals were randomly assigned to either a voice or text chat communication condition. The participants were also randomly assigned to teams and to one of three roles. All members of teams were unfamiliar with each other when they arrived for their sessions.

4.1.1.2 Equipment and Materials. The experiment took place in the CERTT Laboratory configured for the UAS-STE (described earlier). Participants in the text chat condition communicated using the keyboard and a custom-built text communications system designed to log speaker identity and time information. The text communications interface was divided into 3 separate 'modules.' The 'receiver module' alerted participants with a lighted button when a message from another team member was sent. The receiver module also allowed participants to read incoming messages by pressing and holding the F10 key. On releasing the F10 key, the message was then displayed in the 'storage module,' which was comprised of a window that contained previously received messages in a list. Participants were given the ability to scroll through the messages by pressing the F7 and F8 keys. Participants sent messages with the 'transmit module.' To send messages, participants first typed their message in the transmit module

window, selected the recipient using the F3, F4, and F5 keys, and then pressed F1 to send. The interface enabled participants to select multiple recipients. Each message was time stamped with when it was sent (F1 key-presses) and when it was received (F10 key-presses) in order to compute coordination scores (κ) and dynamics. Participants in the Voice Communications condition communicated with each other and the experimenter using David Clark headsets and a custom-built intercom system designed to log speaker identity and time information. The intercom enabled participants to select one or more listeners by pressing push-to-talk buttons.

Custom software (seven applications connected over a local area network) ran the synthetic task and collected values of various parameters that were used as input by performance scoring software. A series of tutorials were designed in PowerPoint for training the three team members. Custom software was also developed to conduct tests on information in PowerPoint tutorials, to collect individual taskwork relatedness ratings, to collect NASA TLX and SART ratings, to administer knowledge questions, and to collect demographic and preference data at the time of debriefing. This report will focus on performance and coordination data.

4.1.1.3 Procedure. The experiment consisted of one 7-hour session. The AVO was located in a separate room adjacent to the other members (DEMPC and PLO). The AVO entered the building through a separate entrance located on the opposite side of the building, and was not allowed to have contact with the other members until debriefing. In the session, the team members were seated at their workstations where they signed a consent form, were given a brief overview of the study and started training on the task.

The number of targets varied from mission to mission in accordance with the introduction of situation awareness roadblocks at set times within each mission. Missions were completed either at the end of a 40-minute interval or when team members believed that the mission goals had been completed. Following each mission, participants were given the opportunity to view their team score, their own individual score, and the individual scores of their teammates. The performance scores were displayed on each participant's computer and shown in comparison to the mean scores achieved by all other teams (or roles) who had participated in the experiment up to that point.

4.1.2 Results

4.1.2.1 Team Performance. Team performance was measured using a composite score based on the result of mission variables including time each individual spent in an alarm state, time each individual spent in a warning state, rate with which critical waypoints were acquired, and the rate with which targets were successfully photographed. Penalty points for each of these components were weighted a priori in accord with importance to the task and subtracted from a maximum score of 1000. Team performance data were collected for each of the five missions.

Team performance was analyzed using a 2 (text, voice) x 4 (mission) mixed ANOVA. Each communication condition (text, voice) had 10 teams. There was a main effect of mission $F(3, 54) = 9.447$, $p < .001$. Teams improved their performance score across the first four missions. There were no significant effects of communication condition, $F(1, 18) = 0.57$, $p < 0.46$, though the voice teams consistently had higher performance scores across all missions than teams in the text chat condition (see Figure 4).

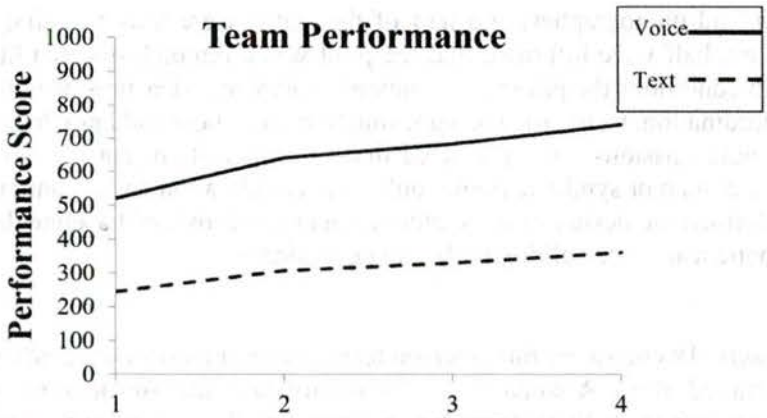


Figure 4. Team performance means for each mission differed over missions, but not condition

LSD pair-wise comparisons showed that team performance improved over the course of the first four missions, with significant gains between the first two missions ($p < 0.05$) and between the second and fourth missions ($p < 0.05$).

4.1.2.2. Coordination. Based on the inherent time costs of using text chat (e.g., typing, noticing a message arrived, etc.), there was a significant time lag between when a message was sent and when it was received (M = 10.5 s for text; 0 s for voice). To determine if there was a difference in coordination score between voice and text chat, a 2 (communication mode) x 4 (lower workload missions) mixed ANOVA was conducted on coordination scores. There was a significant main effect for which text chat had a significantly lower coordination score than voice ($p < 0.05$). This is not to say that the voice condition coordinated "better," but only to say that the two communication conditions coordinated differently. Further, a measure that reveals the stability of team coordination dynamics, the Hurst exponent, was also analyzed to determine if there was a coordination stability difference between communication groups. An independent samples t-test on the average Hurst exponents across teams revealed that text chat teams were, on average, coordinated in a more stable fashion ($M = 0.9527$, $SD = 0.0131$) than voice teams ($M = 0.8988$, $SD = 0.061$), $t(15) = 2.287$, $p < 0.05$.

For the four low workload missions the median of the performance scores was 310 in the chat condition, with 5 teams below the median and 5 teams above the median. A regression analysis on all the teams combined revealed that the linear trend between communication lag and team performance was significant, $F(1, 38) = 9.06$; $p < 0.05$, indicating that as lag decreased, performance increased. Regression analyses also revealed a positive linear relationship between performance score and Kappa in teams performing above the median performance score, $F(1, 13) = 4.46$, $p = 0.055$. Overall these results indicate that text chat results in different coordination patterns than voice chat and that there is a relationship between these patterns and team performance. See Cooke, Myers, & Rajivan (2014) for additional details on this study.

4.2 Experiment 2: Human Expectations of a Synthetic Teammate

In this particular study, we examined how teammate interactions (via text chat) were affected by expectations that the pilot is either a synthetic agent or a human teammate. Three person teams were arranged so that the pilot station and pilot were not visible to the other two teammates

(mission planner and photographer) and half of the teams were informed that the pilot was a synthetic agent and half were informed that the pilot was a remotely-located human teammate. However, in both conditions the pilot was a human participant. Measures of individual and team performance, coordination, team process, team situation awareness and knowledge were collected over four 40-minute missions. We predicted that the expectation that the two teammates are interacting with a human or synthetic pilot would alter coordination and communication patterns. These results informed the design of the synthetic agent and provided baseline data for a Turing-like test of synthetic teammate validity in the next experiment.

4.2.1 Method

4.2.1.1 Participants. Twenty-three three-person teams (team members were unfamiliar with each other) were recruited from Arizona State University and the surrounding community (69 individuals). The teams were divided into two conditions, 10 teams in each condition (synthetic and off-site). These teams participated in one 8-hour session. Due to various hardware or software issues, only 20 teams were able to complete the entire experiment (60 individuals). Of the 60 individuals, 35 were male and 25 were female. Individuals were compensated for their participation by payment of $10 per hour. Individuals were randomly assigned to one of two conditions: off-site AVO or synthetic AVO. The participants were also randomly assigned to teams and to one of three roles role (AVO, PLO, or DEMPC) and were told to work as a team to complete five missions. They made up a team of three specialists who worked together interdependently to take 'good' photos of targets.

4.2.1.2 Equipment and Materials. Participants all used the custom text chat capabilities as in the previous study except the interface was slightly improved to make it easier to use. In addition, the UAS-STE task software was embedded in the new (ONR DURIP funded) CERTT-II hardware as described above.

A series of tutorials designed in PowerPoint were used for training the three team members. Custom software also conducted tests on information in PowerPoint tutorials, collected individual taskwork relatedness ratings, administered knowledge questions, and collected demographic and preference data at the time of debriefing. In addition to software, some mission-support materials (i.e. rules-at-a-glance for each position, two screen shots per station corresponding to that station's computer displays, and examples of good and bad photos for the PLO) were presented on paper at the appropriate workstations. Other paper materials consisted of the consent forms, debriefing forms, and checklists (i.e. set-up, data archiving and skills training).

4.2.1.3 Procedure. Each team participated in one 8-hour session comprised of five missions. Prior to arriving at the session, the three participants were randomly assigned to one of the three task positions: AVO, PLO or DEMPC. The team members retained these positions for the entire study. Team members were seated in locations separated by partitions and did not have face-to-face contact with one another. The AVO was located in a separate room. In the Off-Site condition, the PLO and DEMPC were told that the other team member (i.e., the AVO) was located elsewhere. In the Synthetic Teammate Condition, the PLO and DEMPC were told that they would be working with a synthetic/computer AVO.

The team members were seated at their workstations where they signed a consent form, were given a brief overview of the study and started training on the task. Team members studied three

PowerPoint training modules covering general UAV task knowledge, role-specific and other roles' responsibilities, and how to communicate using the chat system interface at their own pace and were tested with a set of multiple-choice questions at the end of each module. If responses were incorrect, experimenters provided assistance and explanations as to why their answers are incorrect and the reasoning behind the correct answers. Once all team members completed the tutorial, test questions, and communications check, a training mission was started in which experimenters had participants practice the task, checking off skills that are mastered (e.g., the AVO needs to change altitude and airspeed, the PLO needs to take a good photo of a target) until all skills are mastered. Again, the experimenters assisted in cases of difficulty. Training consisted of a total of 1 hour and 30 minutes.

After training, the team started its first 40-minute mission. All missions required the team to take reconnaissance photos of targets. However the number of targets varied from mission to mission in accordance with the introduction of situation awareness roadblocks at set times within each mission (see Table 1). Teams were instructed to obtain as many 'good' photos as they could during the 40 minutes for the mission while avoiding alarms on their interfaces. Missions were completed either at the end of a 40-minute interval or when team members believed that the mission goals had been completed. Immediately after each mission, participants were shown their performance scores. Participants could view their team score, their individual score, and the individual scores of their teammates. The performance scores were displayed on each participant's computer and shown in comparison to the mean scores achieved by all other teams (or roles) who had participated in the experiment up to that point. After the first mission, taskwork knowledge measures and NASA TLX were administered. Once the knowledge measures were completed, teams began the second 40-minute mission followed by the, third, fourth, and fifth missions, the second knowledge session, and concluding with a demographics questionnaire and debriefing.

Table 1. Mission targets counts and roadblocks introduced.

Mission	Number of Targets	Roadblocks
1	11	Communication Glitch PLO to AVO
2	12	New Target ZI
3	11	Target Disguised as Hazard, Communication Glitch DEMPC to PLO
4	11	New Target ZOL, Communication Glitch DEMPC to AVO, New Target SUN
5	20	Dual Communication Glitch PLO to/from AVO

4.2.1.4 ***Measures.*** A performance score was also calculated for each target based on the timely and accurate processing of a target. In addition, a set of behaviors related to team coordination were identified in previous data sets and were noted whenever they occurred in this study. The behaviors are listed in Table 2. Measures included Team performance measured at the mission and target level; Team Process measures of communication (message count), behavioral checklist, and situation awareness. In addition, measures of taskwork and teamwork knowledge were collected as well as workload (i.e., NASA TLX), and demographic questions. All measures except for knowledge measures, NASA TLX, and demographic questions were administered during task performance. Detailed descriptions of the measures are reported in the results section.

Table 2. Individual Behaviors Supportive of Team Coordination

Negative Communication
- Argue – *DEMPC and AVO can argue over the best way to give upcoming waypoint restrictions?*
- Specific to chat conditions
  - Timing – AVO sends text asking for next waypoint just as DEMPC texts the next waypoint info.
  - Lag in response – PLO asks a questions that is not answered until multiple unrelated texts have been posted.

Positive communication
- Help out – *PLO tells DEMPC, "Please give info next target info to AVO."*
- Acknowledge members' speech – *"Roger that."*
- Give praise – *Good job guys!*
- Check with others before implementing a decision – *PLO asks AVO, "I am about to take a pic, are we at 2000 feet?"*
- Clarification – AVO asks DEMPC to clarify what was meant in a previous message.

Repeated Requests
- Same info or action requested two or more times
- PLO asks repeatedly for information needed to take a photo.

Unclear Communications
- Misspellings, ambiguous terms, experimenter cannot understand

General Status Update
- Inform others of current status – AVO tells PLO "I am at 2500 feet now."

Inquiry About Status of Others
- Inquire about current status of others – *DEMPC asks AVO "How are we doing on our heading/fuel etc."*
- Express concern – *DEMPC asks AVO "Are we headed to the next target? We appear to be off course."*

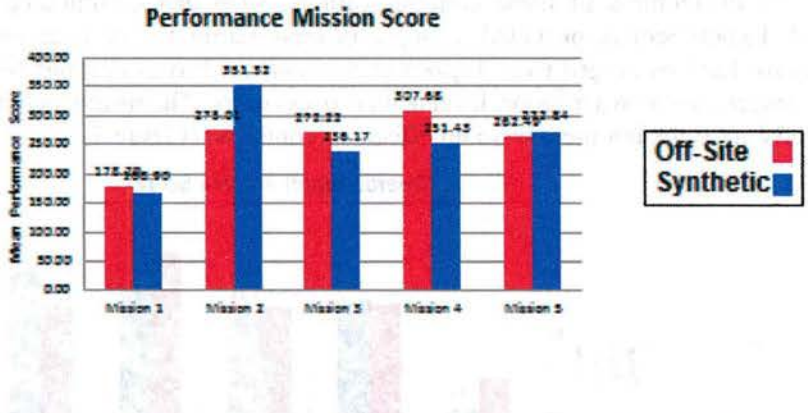
Planning
- Anticipate next steps – *AVO asks DEMPC, "Where are we going after LVN?"*

Suggestions to Others
- Make suggestions to other members – *DEMPC tells AVO to increase speed in route to targets and slow down upon arrival.*

4.2.2 Results

Multiple measures of performance, process, and knowledge were collected in this study. Measures of teamwork knowledge and team situation awareness are not reported due to data collection issues rendering the measure uninteresting. For instance, in the case of the measurement of situation awareness, the roadblocks implemented as communication breakdowns went largely unnoticed given the asynchronous nature of chat. In addition, individual and target level performance scores are not reported here.

4.2.2.1 Team Performance. Team performance, a measure of team effectiveness, was measured after each mission as the weighted composite of team-level mission parameters including time spent in warning or alarm state, number of missed targets, and rate of good target photographs per minute. The rate at which good photos of targets were taken was weighted most heavily. Teams began each mission with a score of 1,000, and points were deducted based on the final values of the mission parameters. This performance score has been validated using other measures of team process and performance. Previous experiments using the UAS task indicate that performance asymptote is consistently reached after four 40-min missions. Results of condition (off-site vs. synthetic) and Mission (1-5 repeated measure) are presented in Figure 6.



Team performance improved across missions (Means for Missions 1 -5 = 173.59, 313.16, 254.19, 279.56, and 267.66, respectively; F(4,72) = 5.709, MSe = 9405.3, p<.0001). The effects of condition (M off-site = 259.14, M synthetic = 256.13, p=.94) and mission by condition interaction(p= .247) were not significant.

Figure 5. Team performance results (Team performance improves, but is not impacted by condition)

4.2.2.2 Team Process: Communication. The three participants of every team communicated with each other via text messages. They used a separate keyboard and touchscreen computer built into their consoles and connected with each other in a virtual network. This mode of communication was the only one available for coordination between the members of each team during each 40-minute mission and was inaccessible at any other time. The chat logger (proprietary software developed in-house) recorded each of these messages with time stamps, recording both the time that it was sent and the time that it was read along with sender and recipient information. Each text message usually consisted of a sentence conveying information about various aspects of the ongoing mission. Sometimes these sentences consisted of cryptic language commonly observed in this mode (e.g. while texting using a cellular device) of communication. In this analysis each individual message was counted regardless of its length or content. The effects of mission ($M_{\text{Missions 1-5}}$ = 97.3, 105.2, 114.6, 106.6, and 102.4 respectively; $F(4, 72) = .82$, $MSe = 988$, $p=.52$), and condition (M off-site = 101.6, M synthetic = 108.7, F(1,18)= .19, p=.518) were not statistically significant. The mission by condition interaction was not significant, $F(4, 72) = .47$, $MSe = 988$, $p=.76$. The number of messages was negatively correlated with team performance, $r(18) = -0.34$.

4.2.2.3 Team Process: Process Ratings. After teams photographed each target, an experimenter rated the quality of the team process behaviors for that target. The experimenters made their ratings by considering three dimensions of team-member interaction at each target: (1) whether the correct information was communicated to the correct team member; (2) the timeliness of those interactions; and (3) quality of communication. The first dimension was based on the degree to which each element of a procedural model of coordination was fulfilled: DEMPC provides the target information to AVO; AVO and PLO negotiate an appropriate airspeed and altitude for that target; PLO provides feedback on the state of the target photo. The second dimension was based on the timing of the interactions relative to the UAV's proximity to the target, such that the relevant interactions were coordinated in time for target processing. The third dimension was based on the

clarity and distinctness of these communications, such that communication events were not repeated. Experimenters provided a single process rating for each target based on the three dimensions. Ratings ranged from 1=*poor* to 5=*excellent*. Process ratings for the first five targets were averaged to obtain a mission-level team process score. The results indicated that team process improved over time, but there were no effects of condition (Figure 7).

Effects of mission (Mission 1 -5 = 1.74, 2.92, 2.81, 3.37, and 3.08, respectively; $F(4, 72) = 9.785$, MSe = 7.718, p<.000) were statistically significant. Neither the mission by condition interaction (p= .856.) nor the effect of condition (p=.887) were statistically significant. Process ratings are positive correlated with mission-level (r(18) =.66, p=.002) performance.

Figure 6. Team process results (Team Process improves over time, but is not impacted by condition).

4.2.2.4 Team Process: Behavioral Checklist. A set of behaviors related to team coordination were identified in previous data sets and were noted by an experimenter whenever they occurred in this study. The behaviors are listed in Table 2. A *2 x 5* (Condition- between Ss x Mission – repeated) Analysis of Variance was conducted for the counts of each of these behaviors in Table 3. Statistically significant effects are reported here for counts of general status update, inquiry about status of others, suggestions to others, and positive communication. All other effects were not significant.

As would be expected as teams develop over time, more proactive updating of an operator's own status increased over time (Figure 8), whereas requests for information about others' status decreased (Figure 9).

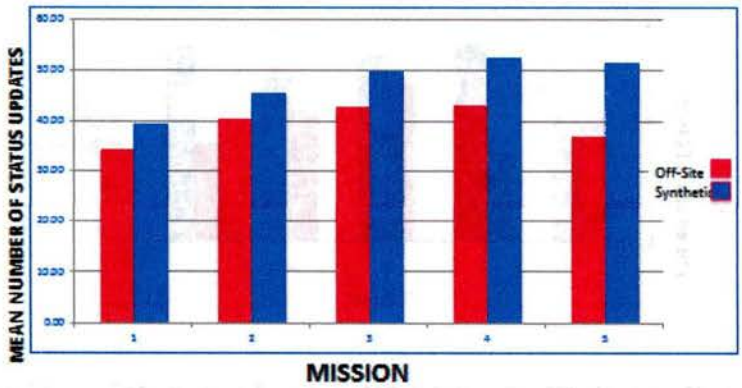


Figure 7. Counts of general status updates (General Status Updates increased for all teams across missions.

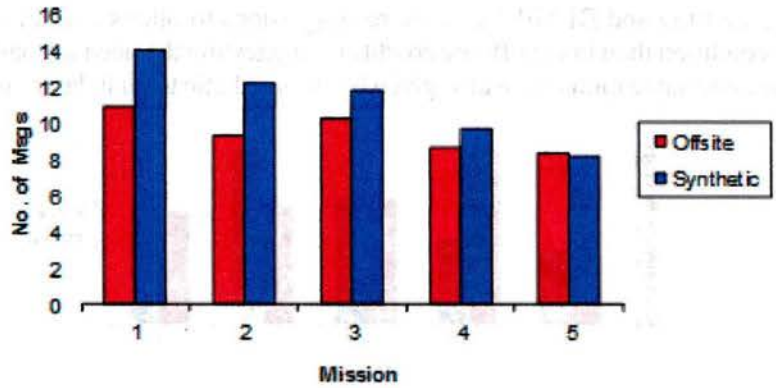
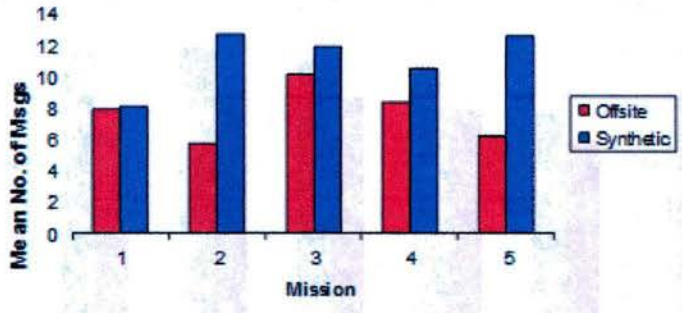


Figure 8. Counts of inquiries of others' status (Inquiries about others status declined over time).

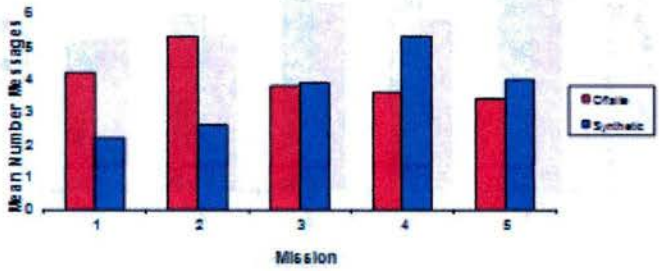
A total of 511 suggestions to others were given in the off-site condition and 696 in the synthetic condition (Figure 10). In each case most of the suggestions (91% - off-site and 95% synthetic) were given by the PLO or DEMPC. Giving suggestions is also correlated with lower target-level performance.



Number of suggestions did not vary over missions (means for Missions 1-5 = 8, 9, 11, 9.5, and 9.5, respectively (p=.36); or condition (M off-site = 7.7, M synthetic = 11.3, p=.45). **The mission x condition interaction F(4, 72) = 2.18, MSe = 5.16, p = .108 was significant at the .11 level. There were more suggestions given in the synthetic condition, especially for Missions 2 and 5.**
The more suggestions given to others, the lower target–level performance r(18)=-.72, p<.0001

Figure 9. Counts of suggestions to others (more suggestions to others were given in synthetic condition).

Interestingly, the PLO and DEMPC gave more suggestions to others (exhibited more control) in the synthetic condition than in the off-site condition, suggesting the need to control the automation. More positive communications were also given by the synthetic team in later missions (Figure 11).



Number of positive communications did not vary over missions (means for Missions 1-5 = 3, 4, 4, 4.5, and 4, respectively (p=.66); or condition (M off-site = 4, M synthetic = 3.6, p=.79). **The mission x condition interaction F(4, 72) = 2.5, MSe = 6.8, p = .05 was significant. Offsite teams had more positive communications in early missions 1 and 2, but synthetic teams had more in later missions 4 and 5.**

Figure 10. Counts of positive communications (Off-site teams had more positive communications in early Missions1 and2, but synthetic teams had more in later Missions4 and5).

4.2.2.5 Team Knowledge. The taskwork knowledge of team members was assessed individually at two points in the experiment: After Mission 1 and after Mission 5. Assessment required participants to provide relatedness ratings of all pairs of 11 terms relevant to the task: altitude, focus, zoom, effective radius, ROZ entry, target, airspeed, shutter speed, fuel, mission time, and photo. All pairs are presented in randomized fashion to participants (one order per pair only). After data collection it was determined that one term was duplicated and one was omitted. Pairs involving either of these terms were dropped from the analysis. Pairwise relatedness estimates were submitted to Pathfinder network scaling (Schvaneveldt, 1990) which created a taskwork

Pathfinder network for each participant. Similarity in terms of the proportion of shared links was calculated for each pair of team members and averaged across the three pairs for each team. Effects of test session, condition, and session by condition interaction were not statistically significant. A comparison of each taskwork network to an empirically-derived taskwork referent resulted in an accuracy score for each team member which was averaged across team members in each condition. As was the case for similarity, the effects of test session, condition, and session by condition interaction were not statistically significant.

4.2.2.6 Workload – NASA TLX. After Missions 1 and 5, a slightly modified version of the NASA TLX was administered to participants individually. The slightly revised NASA Task Load Index (Hart & Staveland, 1988) related to each participant's individual perceptions along these dimensions of workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. The frustration workload was renamed "Emotional workload" in this study. Each dimension was rated along a 1-5 scale across multiple subscales. For instance, the Effort dimension required ratings long a scale anchored on one end by mentally effortless and the other end by mentally difficult. Another sub-scale was anchored as physically effortless to physically difficult. Ratings were all aligned so that a 1 was low and a 5 was high workload. For each dimension the mean of the workload subscales was taken for each team member and means were summed across the six dimensions. A 2x2x3 (Condition (off-Site vs. Synthetic – between Ss) by Mission (1-5, repeated) by Role (AVO, PLO, DEMPC – between Ss) mixed Analysis of Variance was conducted on workload scores. Synthetic teams perceived less workload by the last mission than off-site teams and the DEMPC role was perceived as more difficult than the other two roles. All other effects failed to reach statistical significance (See Figure 12).

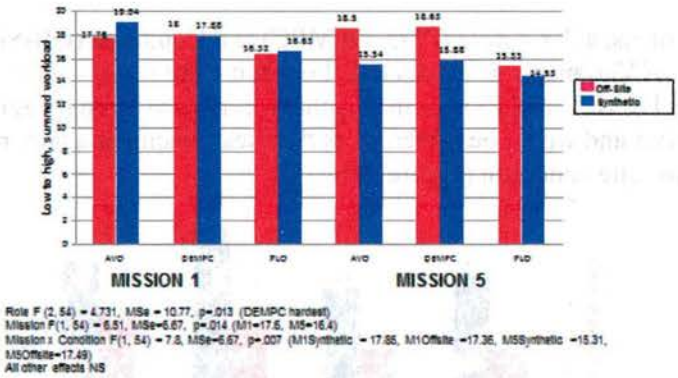


Figure 11. Workload results (Synthetic teams perceived workload by the last mission than off-site teams).

4.2.2.7 Demographics. At the end of the study after the knowledge and NASA TLX questions had been complete, participants were asked a series of questions pertaining to their background and their impression of the experiment. The majority of participants were between 20 and 23 years of age. Of the 60 participants, 35 were male and 25 were female. Teams were composed of both males and females in 63% of the cases. As far as their experience goes, 96% of the participants had played electronic games, 96% had previously participated in an experiment with unmanned vehicle simulators, 76% had taken a course of worked with robotics or remotely controlled vehicles,95% had previously participated in an experiment with robotics or remotely controlled

vehicles (cars, airplanes, boats, etc.). When asked, "At any time during the study, did you suspect that the AVO was a human participant?" 90% of the PLOs and DEMPCs on synthetic teams responded "yes." This value seems high but we suspect that the wording of the question led some to guess that the AVO was not synthetic after all. Because the question is leading we are unable to determine how many participants guessed. On the other hand, we did find significant differences based on the subtle manipulation which involved half of the PLOs and DEMPC's believing that the AVO was synthetic.

Team members were also asked two questions about each of their teammates. Of particular interest are the two questions about the AVO to which the DEMPC and PLO responded:
The AVO on my team was a good member.

a. Strongly Agree
b. Slightly Agree
c. Neutral
d. Slightly disagree
e. Strongly agree

If I were asked to participate in another project like this one, I would like to be with the same AVO.

a. Strongly Agree
b. Slightly Agree
c. Neutral
d. Slightly disagree
e. Strongly agree

For each of these items, a *2 X 2* Role (PLO, DEMPC) and Condition (Off-Site, Synthetic) between subjects Analysis of Variance was conducted. For each there was a significant effect of condition indicating that the DEMPC and PLO in the synthetic condition strongly agreed that the AVO was a good team member and would be preferred as their team member again, more so than DEMPCs and PLOs in the off-site condition (Figure 13).

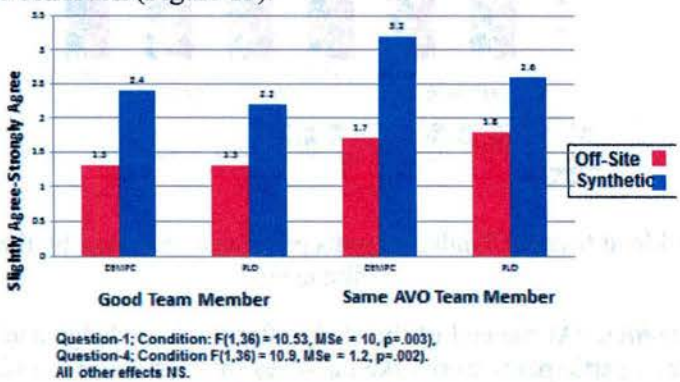


Figure 12. Results of team member judgments (In the synthetic condition participants strongly agree AVO was good team member, also would prefer the same AVO as their team member again).

4.2.2.8 Team Process: Communication Flow

The chat logs for each team and mission were analyzed in terms of message sender and receiver sequence by applying recurrence analysis to the data. We performed a two-factor ANOVA to assess the effect of condition (synthetic vs. offsite) on the recurrence rate of communication. The ANOVA indicated a statistically non-significant effect of condition, $F(1,18) = .001$, $p = .975$. Further the partial η^2 effect size indicated that the interaction explained 0% of the variability for this measure.

We performed another ANOVA to assess the effect of condition (synthetic vs. offsite) on the recurrence rate of communication (determinism. The ANOVA indicated statistically non-significant interaction, $F(1,18) = 1.5$, $p = .230$. Further the partial η^2 effect size indicated that the interaction explained only 7.9% of the variability for this measure which is a small effect by conventional standards (Cohen, 1988).

4.2.2.9 Team Process: Coordination

Coordination is based on the timely sending and receiving of information required for taking good photographs of designated waypoints. A coordination score (κ) is based on the timing and sequence with which key pieces of information are communicated among teammates (Gorman, Amazeen, & Cooke, 2010). The coordination score (κ) is computed as the amount of time from when information I about waypoint w is passed from the DEMPC to the team to when feedback F about taking a good photograph of waypoint w is provided to the team from the PLO. This is then divided by the amount of time from when the PLO and AVO negotiate N UAV flight dynamics for waypoint w up to when the PLO provides feedback F that a good photograph was taken for the waypoint w.

$$\kappa = \frac{F_w - I_w}{F_w - N_w} \tag{1}$$

The κ scores were computed for each target processed by each team in sequence across all five missions. Because not all teams processed targets in the same order, sequences of targets were not identical. Lyapunov exponents were calculated for each of the sequences. There was no significant difference found between the synthetic and offsite condition, though there was a suggestion that with additional statistical power a reliable difference may be attained.

4.2.3 Discussion

These results indicated that the subtle manipulation of telling half of the teams that the human AVO was synthetic had some interesting effects. These effects speak to human expectations regarding synthetic teammates or more generally, automation. The PLO and DEMPC in the synthetic condition gave more suggestions than those in the off-site condition suggesting that the two teammates had a greater need for control when they worked with a synthetic teammate than when they worked with a human teammate. Individuals on synthetic teams also provided more positive feedback in later missions than off-site teams and the PLO and DEMPC of synthetic teams perceived less workload than off-site PLOs and DEMPCs on those same later missions. The PLO and DEMPCs of synthetic teams also liked working with the AVO more so than those on off-site teams. Therefore when people thought they were working with a synthetic AVO they exerted more control, but perceived less workload and generally liked the synthetic teammate. They seemed impressed with the automation. The idea of controlling automation is interesting in the

context of human-automation teaming. It may be difficult or unnatural for humans to give up control to a nonhuman teammate.

These results also speak to coordination behaviors of good teams. As teams develop through the course of the experiment, they exhibit less pulling and more pushing of information. The following task of integrating the synthetic teammate into the new UAS-STE occurred in parallel with the first two experiments. Once integration was complete we were able to conduct the evaluation study that follows.

4.3 Integration of Synthetic Teammate into New UAS-STE

Connectivity between the synthetic teammate at AFRL-Dayton and the CERTT Lab in Mesa, AZ created several challenges for this project. Preliminary connectivity was established between the PALM Lab at WPAFB and CERI in Mesa, AZ, however, due to issues with firewalls and difficult access to AFRL's internet, it was decided that we would physically relocate the synthetic teammate software/hardware to CERI in Mesa, AZ.

Also, in the course of connectivity and integration efforts, our team ran into difficulty with the CERTT-I task software running on the updated CERTT II hardware. (CERTT-I software was a RAPID-based code used in the nine previous CERTT UAV-STE studies in the lab including the baseline studies described above. The recent DURIP funding provided new task software and hardware for the lab – CERTT-II – however, the synthetic teammate was developed in accord with the CERTT–I task software and testing needed to be conducted on the same task software.) RAPID is no longer in existence and was causing problems in the Synthetic Teammate integration efforts. Software modifications and upgrades to the task software became impossible.

Specific Issues included:

- RAPID used DDE communications for inter-computer communications. Windows 7 64-bit does not support DDE communications
- Data was stored from the RAPID task software into single flat text files.
- Enabling communications between the Synthetic AVO and the overall system was proving to be overly cumbersome
- Specific metrics that are currently needed were not being recorded in the RAPID system
- System was susceptible to communications glitches which would require a mission restart

A decision was made to re-write the CERTT–I Task software using Visual Studio.Net. The CERTT-I task software was completely re-written. Features of this new system include:

- Compatible with latest operating systems – we are using Win 7 – 64 bit machines
- Complete replication of the original CERTT user interface screens
- Robust TCP-IP communications protocols for inter-machine communications
- The ability to use either human or synthetic teammates in any of the three task roles
- Performance and metrics data recording to a SQL database.
- Playback capability
- Automatic loading of different scenarios for each mission
- All scenario and startup data contained in a SQL database
- Scenario editing/authoring software module developed
- Advanced mission control system and interface

- Automatic recovery from a communications failure without mission disruption
- An interface control document was developed for the remote (synthetic clients)
- Ability to pause and resume a mission if needed
- Tight integration of Coordination Logger and other external measures – all logged to single SQL database
- This is now a flexible platform for future human – automation research experiments

Once the synthetic teammate was in place in Mesa, AZ and the new CERTT-II software was installed and tested, we ran iterative trials with project researchers as participants to test and iteratively refine the synthetic teammate and especially, its interface with the task and two human participants. One of the things that we ended up doing was providing participants with very specific instructions for interacting with the AVO (See Appendix A). This was to avoid a few issues that the synthetic teammate had with comprehending the wide range of natural text language that participants would use. The human participants were provided with a template with which to converse with the synthetic teammate. Although more time could have been spent perfecting the synthetic teammate's natural language comprehension, we were anxious to finesse that process so that we could learn about the deeper issues concerning team interaction.

4.4 Experiment 3: Synthetic Teammate Evaluation

Once the synthetic teammate had been adequately tested in the CERTT-II testbed, we conducted an experiment to: 1) evaluate the synthetic teammate's performance as a member of a 3-agent team (in the pilot role), 2) evaluate the human-synthetic team performance in comparison to all human teams, 3) understand how team process differs between all human and human-synthetic teams and how this impacts performance, and 4) compare the human-synthetic teams and all human control teams to a team with an pilot that is experienced in pushing and pulling information across the team. The first two objectives provided a Turing-like test of synthetic teammate validity. The third objective informed scientific understanding of effective teamwork. The last objective helped establish the upper boundaries of team performance in this task that might be achieved with a highly effective synthetic pilot.

4.4.1 Methods

4.4.1.1 Experimental Task and Procedure. Like the previous experiments, this experiment was conducted in the context of the CERTT UAS-STE (Cognitive Engineering Research on Team Tasks Unmanned Aerial System - Synthetic Task Environment (Cooke & Shope, 2004, 2005). A single UAS-STE mission consists of 11-12 targets and lasts a maximum of 40 minutes; each team performs five 40-minute missions (see Figure 14). Measures are taken in the context of each mission and in some cases (knowledge, demographics, workload) apart from missions.

Figure 13. CERTT UAS-STE team roles and task.

4.4.1.2 Experimental Design and Hypotheses. We tested three types of teams (conditions), each with three teammates, in a 5-mission CERTT-UAS-STE task: 1) the synthetic condition - the AVO role is given to the synthetic teammate; 2) the control condition - the AVO was an inexperienced human participant just like the other participants (PLO and DEMPC); and 3) the experimenter condition–one of the experimenters served as an expert AVO, pushing and pulling information across the teammates in a scripted manner (see Figure 15). Specifically, in experimenter condition, the AVO asked questions of other team members in a scripted fashion to ensure timely and adaptive passing of information at target waypoints.

We hypothesized that teams with a synthetic teammate would demonstrate poorer team performance and different team process than the control and experimenter conditions due to subtle teamwork deficits in the synthetic teammate. We further hypothesized that control teams would display poorer performance than the experimenter teams due to less efficient coordination.

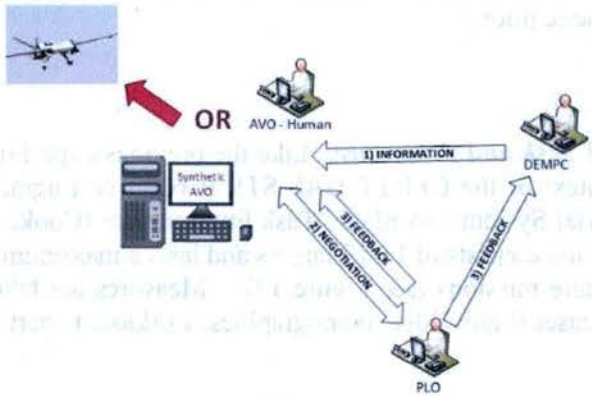


Figure 14. Two conditions used a human AVO (experienced or inexperienced) and one condition used the synthetic AVO.

4.4.1.3 Measures. Measures of individual and team performance process and knowledge were collected and we report on results associated with those indicated in Table 3. We do not report results of demographics, coordination, communication, knowledge, and workload here.

Table 3. AF10 Measures and Result Highlights

Measures	Result Highlights
Team Performance	Synthetic=Control<Experimenter
AVO Performance	Synthetic<Control<Experimenter
Target Processing Efficiency	Synthetic<Control<Experimenter
Process Ratings	Synthetic=Control<Experimenter
Team Verbal Behavior	Synthetic teams pull more than the push
Team Situation Awareness	Synthetic=Control<Experimenter

4.4.1.4 Participants. Thirty teams (70 participants) were recruited for participation from Arizona State University, and completed the experiment. For the experimenter and synthetic teammate conditions, two participants per team were recruited for the PLO and the DEMPC roles, and the role of AVO was played by either a trained confederate (experimenter condition) or synthetic teammate (synthetic condition). Participation required normal or corrected-to-normal vision and fluency in English. Participants ranged in age from 18 to 38 (M_{age}=23.676, SD_{age}= 3.294) and 60 were male and 10 were female. The teams were composed of undergraduate and graduate students. Each team participated in one seven-hour session, and each individual was compensated for participation by payment of $10 per hour.

4.4.2 Results

4.4.2.1 Team Performance. Teams begin each mission with a score of 1,000, and points were deducted based on the final values of the mission parameters. We performed a 3 (Condition) x 5 (Missions) mixed Analysis of Variance (ANOVA) design to determine whether the three conditions -synthetic, control, and experimenter- differed with respect to improvement of their mission scores over time (i.e., repeated measure ~ five missions). In the synthetic teammate condition, four mission level scores were missing because of technical difficulties of the synthetic teammate.[1] During these missions, the synthetic teammate did not interact with human team members even if the human team members communicated with the synthetic teammate in a well-structured and timely manner. This situation adversely affected these teams' individual and team scores. Thus, in order to handle the missing data, multiple imputation was used before the analysis.

The ANOVA indicates that the condition main effect, $F(2, 27) = 11.304$, $p < .05$ (with a large effect size, $\eta^2 = .45$), and the repeated measures (i.e., mission) main effects, $F(3.141, 84.818) = 3.683$, $p < .05$ (with a medium effect size, $\eta^2 = .120$) were statistically significant. However, the condition by mission interaction effect was not significant, $F(8, 108) = 1.498$, $p = .166$.Only experimenter teams demonstrated a learning effect in that team performance improved across the missions (Mission 1 to Mission 4, $p < .05$). On the other hand, the synthetic teams' performance decreased (from Mission 1 to Mission 5, $p < .05$), and the teams in the control condition showed no significant change across the missions.

[1] There are four missing data points, because of technical difficulties with the synthetic teammate: Team 3-Mission 5, Team 16-Mission 1, Team 17-Mission 4, and Team 28-Mission 7.

In terms of condition level comparisons, the experimenter teams (*MExp*= 397.574, *SDExp.* = 94.893, $p < .05$) performed better than synthetic and control teams, (*MSynthetic* = 293.130, *SDSynthetic* = 93.559, *MControl*= 248.645, *SDControl* = 95.34, $p < .05$). The synthetic teams did not perform significantly better than control teams, $p = .178$; Figure 16).

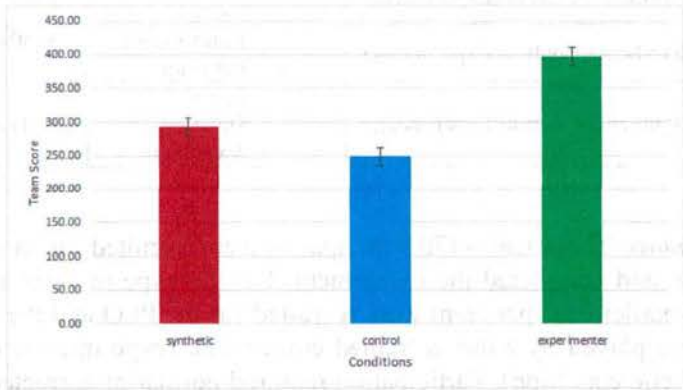


Figure 15. Team Performance across the Conditions (**Synthetic=Control<Experimenter**). Error bars provide the standard error of the mean

4.4.2.2 AVO Performance. There are separate outcome-based individual performance scores as well. In this experiment, because the pilot (AVO) is the main manipulation, we have only reported the AVO performance score which was obtained from the summation of four penalty scores (i.e., alarm and warning penalties, route sequence penalty, and course deviation penalty) subtracted from 1000. We performed a 3x5 mixed Analysis of Variance on the AVO score. The ANOVA analysis indicates that the following three effects were statistically significant: the condition main effect, $F(2, 27) = 16.858$, $p < .001$, with a large effect size, $\eta^2 = .555$; the repeated measures (i.e., mission) main effect, $F(1.563, 42.209) = 6.413$, $p < .05$, with a small effect size, $\eta^2 = .192$; and the interaction effect between condition and mission, $F(8, 108) = 3.656$, $p < .05$, with a medium effect size, $\eta^2 = .213$. According to the Mission simple effects, though the synthetic teammate performance decreased across the missions (from Missions 1 to Mission 4, $p < .05$), the AVO performance in the control condition increased across the missions (from Mission 1 to Mission 4, $p < .05$). The AVO performance score in the experimenter condition was stable across the missions. The synthetic teammate performed significantly worse than the control AVO (*MSynthetic* = 607.820, *SDSynthetic* =288.565, *MControl*= 764.58, *SDControl* = 94.838, $p < .001$), and the AVO in the control condition performed significantly worse than the AVO in the experimenter condition (*MExp*= 853.576, *SDExp.* = 67.674, $p < .05$; Figure 17).

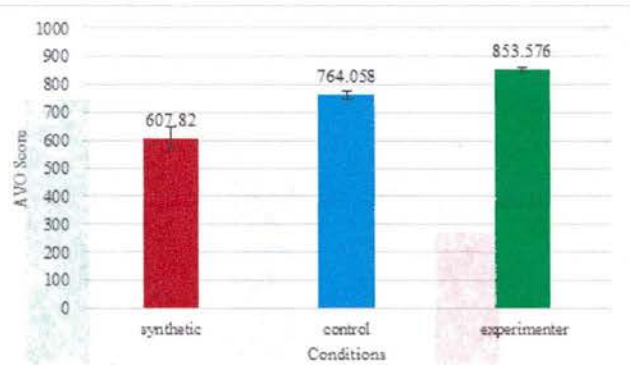


Figure 16. AVO Performance across the Conditions (**Synthetic <Control<Experimenter**). Error bars provide the standard error of the mean.

4.4.2.3 Target Processing Efficiency. Target Processing Efficiency took into account the time spent inside a target waypoint to get a good photo (higher scores equate to more efficiency). The target level team performance score was analyzed via a basic model with a three level nested mixed Analysis of Variance (ANOVA) was conducted with condition as a between teams manipulation and mission and target as within teams: target (level 1), within mission (level 2), and between condition (level 3). According to the mixed ANOVA results, there was a significant condition main effect $F(2, 32.243)= 10.853$, $p < .001$, with a medium effect size $\eta^2= .402$. However, there was no mission main effect, $F(4, 309.320)=2.129$, $p =.077$, nor a condition by mission interaction effect, $F(8,195.608)=1.273$, $p =.259$.

At the target level of team performance, a learning effect was found for the synthetic condition from Mission 6 to Mission 8 ($M_3 = 776.914$, $SD_3 =$, $M_9 = 868.158$, $SD_9 =$, $p < .05$), and for the control condition from Mission 2 to 5 ($M_2 = 841.612$, $SD_2 =$, $M_5 = 901.706$, $SD_5 =$, $p < .05$). The synthetic teams' performance decreased from Mission 2 to 3 ($M_2 = 846.330$, $SD_2 =$, $M_3 = 776.914$, $SD_3 =$, $p < .05$). This indicates that in the beginning of the experiment, the synthetic teams did not perform well as same as control and experimenter teams. The experimenter teams demonstrated a non-significant, but continuous increase in performance from Mission 1 to Mission 5 ($M_1 =928.003$, $SD_1 =$, $M_5 = 953.598$, $SD_5 =$, $p = .184$).

According to the condition level comparisons, the synthetic teams performed significantly worse than the control and experimenter teams, $M_{Synthetic} = 820.087$, $SD_{Synthetic} = 158.413$, $M_{Control} = 866.173$, $SD_{Control} = 119.548$, $M_{Exp} = 939.184$, $SD_{Exp} = 58.026$, $p < .001$ and the experimenter teams performed significantly better than control and synthetic teams ($p < .05$; Figure 18).

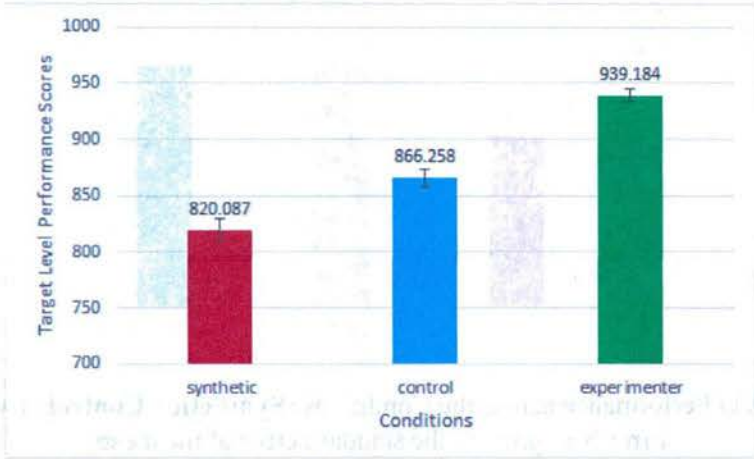


Figure 17. Target Processing Efficiency for the Control, Synthetic, and Experimenter between-subjects conditions (**Synthetic < Control < Experimenter**). Error bars provide the standard error of the mean.

4.4.2.4 Summary: Team Performance. Taken together the results of the outcome-based and target processing efficiency scores support our hypotheses that the synthetic AVO does not perform as well as an inexperienced AVO who does not perform as well as an experienced AVO. At the team level, synthetic teams perform as well as control teams on the outcome-based measure of team performance, though the experimenter outperforms both. However, the synthetic teams so not process targets as efficiently as control teams who are less efficient than experimenter teams. Processing efficiency requires efficient team interaction. We turn next to our process measures to better understand the synthetic team deficits.

4.4.2.5 Team Process: Process Ratings. During the experiment, two experimenters rated the quality of the team process behaviors for the current target, after it was photographed by the teams. The experimenters considered three dimensions of team member interaction at each target: 1*) Coordination*: communicating with the correct team member about the correct information based on the optimal coordination sequence, 2) *Timeliness of the Coordination* of the target based on the team's ability to coordinate through relevant interaction in time to process the target (requires looking at when the interactions occurred and the UAV's relative position to the target); and 3) *Quality of Communication*, based on how clear and unique the communications are (ideally, minimizing the need for repetition). Two of the experimenters rated process at each target independently based on these three dimensions, with ratings ranging from 1= *poor* and 5 = *excellent*. Therefore, weighted Cohen's κ was run to determine if there was an agreement between two experimenters' observations on recording the team process ratings. There was fair agreement between the two experimenters' observation, $\kappa = .362$ (95% CI, .319 to .405), $p < .001$. In order to measure target level process rating, a basic model with three levels nested mixed Analysis of Variance (ANOVA) was conducted with condition as a between teams manipulation and mission and target within: target (level 1), within mission (level 2), and between condition (level 3).

According to the results, there was a significant condition main effect $F(2, 35.644) = 35.825$, $p < .001$, with a large effect size $\eta^2 = .668$, and mission main effect, $F(4, 300.519) = 2.848$, $p < .05$, with a small effect size $\eta^2 = .037$ but no condition x mission interaction effect, $F(8,200.317) = .876$,

$p = .538$. The pairwise dependent t tests (for conditions) indicates that though teams in the synthetic condition ($M_{synthetic} = 2.246$, $SD_{synthetic} = .416$) had significantly lower process ratings than those in the experimenter condition ($M_{exp} = 3.171$, $SD_{exp} = .619$), the experimenter condition ($M_{exp} = 3.171$, $SD_{exp} = .619$) had significantly higher team process ratings than control ($M_{control} = 2.255$, $SD_{control} = .694$) conditions ($p < .001$). However, comparisons between synthetic ($M = 2.246$, $SD = .416$) and control ($M = 2.255$, $SD = .694$) were not statistically significant ($p = .911$), mirroring previous team performance findings (Figure 19).

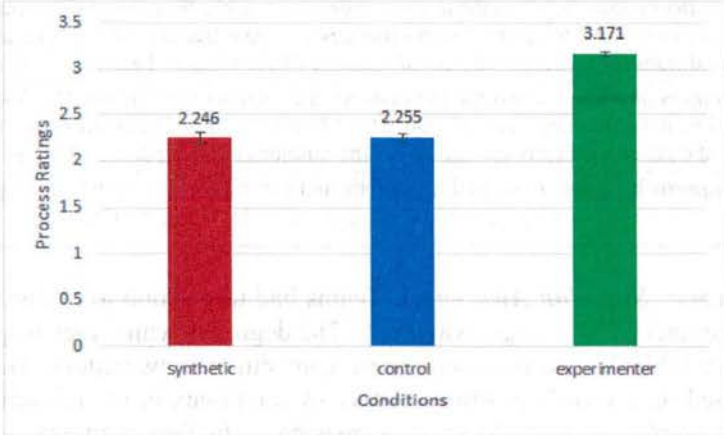


Figure 18. Target Level Team process ratings for the Control, Synthetic, and Experimenter between-subjects conditions (**Synthetic = Control < Experimenter**). Error bars provide the standard error of the mean.

4.4.2.6 Team Process: Team Verbal Behaviors. Eight verbal behaviors were identified from previous CERTT-UAS-STE data as being associated with team effectiveness (Table 4). Instances of these behaviors were tagged by two experimenters. Cohen's κ was run to determine if there was an agreement between two experimenters' observations on recording the teams' verbal behaviors, and it was found that there was substantial agreement between the two experimenters' observation, $\kappa = .774$ (95% CI, .754 to .794), $p < .001$.

Table 4. Team verbal behaviors

Behaviors	Description
General Status Updates	informing other team members about current status
Repeated Requests	requesting the same information or action from other team member(s)
Inquiry about Status of Others	inquiring about current status of others, and expressing concerns
Suggestions	making suggestions to the other team members
Planning Ahead	anticipating next steps and creating rules for future encounters
Positive Communication	helping out team members by providing information and acknowledgement of member's speech
Negative Communication	argument among the team members due to conflicting goals or incorrect destination
Unclear Communications	sending information with misspellings and ambiguous terms which experimenters cannot understand

In order to analyze the verbal behaviors, we performed a 3 (condition) x 5 (mission) repeated measures Multivariate ANOVA on eight team verbal behaviors for each role. In the interest of space, only statistically significant differences between the synthetic and other conditions are summarized in Table 5 below.

As can be seen in Table 5, human-synthetic teams gave fewer general status updates and had more repeated requests compared to other teams. This paints a general picture of teams that are doing more pulling than pushing of information. Also human-synthetic teams did not decrease inquiries to others or this pulling behavior over time as higher performing teams did suggesting that more implicit coordination patterns were not developing.

Table 5. Verbal behavior differences between synthetic and other teams.

General status updates between the team members were more frequent in the experimenter condition over time while the general status updates were less frequent in synthetic and control conditions ($M_{experimenter}$= 25.520, $M_{synthetic}$= 14.900, $M_{control}$= 19.270, $p < .05$).
Inquiries to others decreased across the missions for the control (M_1= 14.200, M_5= 7.400, $p < .05$) and the experimenter conditions (M_1= 17.900, M_5= 12.650, $p < .05$). The synthetic teams' inquiries to others did not significantly change across the missions (M_1= 5.50, M_5= 5.80, p = .892).
Repeated requests happened more in the synthetic and control conditions than the experimenter condition ($M_{synthetic,}$= .710, $M_{control}$= .470, $M_{experimenter}$=. 410).

4.4.2.7 Team Process: Situation Awareness. Teams had to respond to occasional "roadblocks" such as the introduction of a new target waypoint. The degree to which they responded accurately as a team to these roadblocks was one measure of team situation awareness. Because roadblocks were triggered based on a team's position relative to waypoints in the mission, each team may trigger a different number of roadblocks in a mission. In fact, synthetic and control teams triggered fewer roadblocks than the experimenter teams.

In order to analyze completion of the roadblocks after being triggered, first, repeated measures logistic regression was run to examine the number of roadblocks completed. The results indicates that the main effect of the *Condition* (χ^2 (2) = 17.778, $p < .001$) was statistically significant. However, because there were no significant repeated measure (i.e., mission) effect (χ^2 (4) = 7.947, p =.094) and mission x condition interaction (χ^2 (8) = 7.043, p = .532), these two effects were dropped and binomial logistic regression was performed three times to ascertain the effect of condition on the likelihood that teams overcame the roadblocks. The logistic regression model was statistically significant, χ^2 (2) = 28.997, $p < .001$. The model explained 17.5% (Nagelkerke R^2) of the variance in completing the roadblocks, and correctly classified 71% of cases.

According to the findings, though experimenter teams were 4.169 times more likely to overcome the roadblocks than control teams, χ^2 (1) = 15.341, $p < .001$), synthetic teams were 0.144 times less likely to overcome the roadblocks than experimenter teams, (χ^2 (1) = 22.213, $p < .001$. Even though the synthetic teams overcame fewer roadblocks than the control teams, the difference between synthetic and control was not significant, χ^2 (1) = 1.689, p = .194 (Figure 20). Thus, the team situation awareness data mirrors the team performance data.

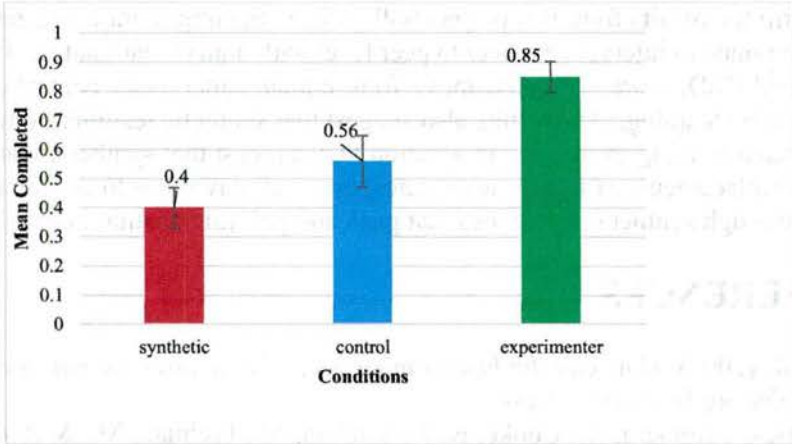


Figure 19. Estimated means of proportion of overcoming the roadblocks after triggered (**Synthetic = Control <Experimenter**). Error bars provide the standard error of the mean.

4.4.3 Discussion

Data reported here provide a consistent picture of the Synthetic teammate performance in comparison to Control and Experimenter teams. Findings indicate:

- Synthetic AVOs perform more poorly than control AVOs
- Synthetic teams perform as well at the mission level as control (all human) teams, but process targets less efficiently
- Synthetic teams demonstrate interaction patterns corresponding to more pulling of information than pushing with little change over time. This seems to indicate immaturity in terms of development of team coordination. Future coordination analysis will investigate this more deeply.
- The stellar team performance across the board of the Experimenter teams demonstrates what can be achieved by inserting an "Expert" synthetic teammate into a team training exercise. Training happens implicitly through the deliberate and timely pushing and pulling of information by the "expert" agent.

5.0 DISCUSSION

The development and evaluation of an autonomous agent that acts as a full-fledged teammate represents an important and first-of-a-kind achievement in human-autonomy teaming.

Through this effort we have developed an understanding about how humans interact with a synthetic teammate and how the synthetic agent impacts team performance and team coordination (e.g., Demir & Cooke, 2014). In addition we have learned how human expectations of autonomy can affect interactions with autonomy and how mode of communication can affect team coordination (Cooke, Myers, & Rajivan, 2014). Most importantly, together with AFRL we have developed and evaluated a fully autonomous teammate that is capable of interacting with two human teammates in a complex RPAS control environment (i.e., the CERTT Lab), achieving levels of performance comparable to all-human teams under nominal conditions.

In the long term the results from this project will indicate the important characteristics needed of a synthetic teammate to interact on a peer to peer level with human teammates. This is significant as the Navy and DoD, more generally, move from human supervisory control of automation to human-automation teaming. The results also suggest that synthetic teammates may act as a force multiplier in team training exercises. In addition they suggest that synthetic teammates may not only serve as replacements of human teammates, but also may serve to accelerate training team coordination through synthetic teammates that push and pull information in an efficient manner.

6.0 REFERENCES

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford; New York: Oxford University Press.

Ball, J., Myers, C., Heiberg, A., Cooke, N. J., Matessa, M., Freiman, M., & Rodgers, S. (2010). The synthetic teammate project. *Computational and Mathematical Organization Theory*, *16*(3), 271–299. http://doi.org/10.1007/s10588-010-9065-3

Baltes, B. B., Dickson, M. W., Sherman, M. P., Bauer, C. C., & LaGanke, J. S. (2002). Computer-Mediated Communication and Group Decision Making: A Meta-Analysis. *Organizational Behavior and Human Decision Processes*, *87*(1), 156–179. http://doi.org/10.1006/obhd.2001.2961

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

Cooke, N. J., & Gorman, J. C. (2009). Interaction-Based Measures of Cognitive Systems. *Journal of Cognitive Engineering and Decision Making*, *3*(1), 27–46. http://doi.org/10.1518/155534309X433302

Cooke, N. J., Gorman, J. C., Duran, J. L., & Taylor, A. R. (2007). Team cognition in experienced command-and-control teams. *Journal of Experimental Psychology: Applied*, *13*(3), 146–157. http://doi.org/http://dx.doi.org.ezproxy1.lib.asu.edu/10.1037/1076-898X.13.3.146

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive Team Cognition. *Cognitive Science*, *37*(2), 255–285. http://doi.org/10.1111/cogs.12009

Cooke, N. J., Gorman, J., Pedersen, H., Winner, J., Duran, J., Taylor, A., … Rowe, L. (2007). *Acquisition and Retention of Team Coordination in Command-and-Control*. Retrieved from http://www.dtic.mil/docs/citations/ADA475567

Cooke, N. J., Myers, C. W., & Rajivan, P. (2014). In D. M. A. Vidulich, P. S. Tsang, & J. M. Flach (Eds.), *Advances in Aviation Psychology: Volume 1* (pp. 141–157). Ashgate Publishing, Ltd.

Cooke, N. J., Rivera, K., Shope, S. M., & Caukwell, S. (1999). A Synthetic Task Environment for Team Cognition Research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*(3), 303–308. http://doi.org/10.1177/154193129904300337

Cooke, N. J., & Shope, S. M. (2004). Designing a Synthetic Task Environment. In L. R. E. Schiflett, E. Salas, & M. D. Coovert (Eds.), *Scaled Worlds: Development, Validation, and Application* (pp. 263–278). Surrey, England: Ashgate Publishing. Retrieved from http://www.cerici.org/documents/Publications/scaled%20worlds%20paper3.pdf

Cooke, N. J., & Shope, S. M. (2005). Synthetic Task Environments for Teams: CERTT's UAV-STE. In N. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. Hendrick (Eds.), *Handbook of Human Factors and Ergonomics Methods* (pp. 41–46). Boca Raton: FL: CRC Press.

Retrieved from http://www.certt.com/publications/Synthetic%20Task%20Environments%20for%20Teams.pdf

Demir, M., & Cooke, N. J. (2014). Human Teaming Changes Driven by Expectations of a Synthetic Teammate. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 16–20. http://doi.org/10.1177/1541931214581004

Gorman, J. C., Amazeen, P. G., & Cooke, N. J. (2010). Team coordination dynamics. *Nonlinear Dynamics, Psychology, and Life Sciences*, *14*(3), 265–289.

Gorman, J. C., & Cooke, N. J. (2011). Changes in team cognition after a retention interval: The benefits of mixing it up. *Journal of Experimental Psychology: Applied*, *17*(4), 303–319. http://doi.org/http://dx.doi.org.ezproxy1.lib.asu.edu/10.1037/a0025149

Gorman, J. C., Cooke, N. J., & Amazeen, P. G. (2010). Training Adaptive Teams. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(2), 295–307. http://doi.org/10.1177/0018720810371689

Gorman, J. C., Cooke, N. J., Pedersen, H. K., Winner, J., Andrews, D., & Amazeen, P. G. (2006). Changes in Team Composition after a Break: Building Adapative Command-and-Control Teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(3), 487–491. http://doi.org/10.1177/154193120605000358

Gorman, J. C., Cooke, N. J., Pederson, H. K., Olena, O. C., & DeJoode, J. A. (2005). Coordinated Awareness of Situation by Teams (CAST): Measuring Team Situation Awareness of a Communication Glitch. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(3), 274–277. http://doi.org/10.1177/154193 120504900313

Gorman, J. C., Cooke, N. J., & Winner, J. L. (2006). Measuring team situation awareness in decentralized command and control environments. *Ergonomics*, *49*(12–13), 1312–1325. http://doi.org/10.1080/00140130600612788

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Mashkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland Press.

Rodgers, S. M., Myers, C. W., Ball, J., & Freiman, M. D. (2013). Toward a situation model in a cognitive architecture. *Computational and Mathematical Organization Theory*, *19*(3), 313–345. http://doi.org/10.1007/s10588-012-9134-x

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: studies in knowledge organizations*. Norwood, N.J.: Ablex Pub. Corp.

Weeks, G. D., Kelly, M. J., & Chapanis, A. (1974). Studies in interactive communication: V. Cooperative problem solving by skilled and unskilled typists in a teletypewriter mode. *Journal of Applied Psychology*, *59*(6), 665–674. http://doi.org/http://dx.doi.org.ezproxy1.lib.asu.edu/10.1037/h0037499

7.0 TRANSITIONS

- CERTT Chat Data collected under this effort has been shared with AFRL's PALM Lab to serve as a language corpus for synthetic teammate development and for use in the evaluation of and iteration on synthetic teammate models (Chris Myers; christopher.myers.29@us.af.mil).
- CERTT Chat data has also been shared with Dr. Jamie Gorman of Georgia Tech who is applying new dynamical analyses related to fractal patterns to these data (Jamie Gorman; Jamie.Gorman@psych.gatech.edu).

8.0 COLLABORATIONS

- **CRADA:** The Synthetic Teammate project benefits from a cooperative research and development agreement (CRADA) between members of the Performance and Learning Models (PALM) team of the Air Force Research Laboratory, 711th Human Performance Wing, Human Effectiveness Directorate, Warfighter Readiness Research Division (AFRL/RHAC) and the Cognitive Engineering Research Institute (CERI).
- **MOU:** CERI and Arizona State University have signed a Memorandum of Understanding that permits the sharing of students and equipment. Cooke has been collaborating with Subbarao Kambhampati, also at ASU, on a project on
- **Human-Robot Teaming:** Planning for Peer-to-Peer Human Robot Teaming in Open Worlds. Again the issue is human-robot teaming.
- **Sandia Research:** Sandia Research has started commercializing the CERTT-II Unmanned Aerial System-Synthetic Task Environment and it now resides in Joshua Woolley's MD, PhD, BAND Lab http://woolleylab.ucsf.edu/ run as a cooperation between the CA and University of California at San Francisco.
- **DURIP:** The project was also facilitated by the ONR Defense University Research Instrumentation Program (DURIP) grant which funded CERTT-II, a more flexible, stable testbed platform.
- **Prior Support:** Earlier work on the synthetic teammate project was funded by the Air Force Office of Scientific Research (AFOSR) and the Air Force Research Lab (AFRL). Metrics and findings that are leveraged have received support from AFOSR, AFRL, and ONR.

9.0 PERSONNAL SUPPORT

- **Principal investigator:** Nancy Cooke
- **Graduate Students:** Cade Bartlett (MS), Jayanta Das (PhD), Mustafa Demir (PhD), Bryant Foster (MS), Lourdes Reyes (MS), Mistey Taggart (MS)
- **Undergraduates:** Ashley Knobloch, Ivonne Murray, Sara Zipp
- **Summer high school students:** Savleen Kaur, Nikil Patel
- **Sandia Research Corporation subcontractors:** Mike Dinan, Paul Jorgenson, Steven Shope, Amanda Taylor, Kyle Uithoven

10.0 PUBLICATIONS

10.1 Journal Publications

Cooke, N. J. (2015). Team cognition as interaction. *Current Directions in Psychological Science, 34*, 415-419.

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J.L. (2013). Interactive Team Cognition, *Cognitive Science, 37*, 255-285, DOI: 10.1111/cogs.12009.

Gorman, J. C., Hessler, E. E., Amazeen, P. G., Cooke, N. J., & Shope, S. M. (2012). Dynamical analysis in real time: Detecting perturbations to team communication. *Ergonomics, 55,* 825-839. http://dx.doi.org/10.1080/00140139.2012.679317

10.2 Book Chapters

Cooke, N. J., Myers, C. W., & Rajivan, P. (2014). *Implications of text chat for crew communication and coordination.* In M. Vidulich, P. Tsang, and J. Flach (Eds.). *Advances in Aviation Psychology: Volume I, (pp. 141-158).* Surrey, England: Ashgate.

Cooke, N. J., & Gorman, J. C. (2013). Microworld Experimentation with teams. In A. Kirlik & J. D. Lee (Eds.), *The Oxford Handbook on Cognitive Engineering, (pp. 327-335) NY: Oxford Press.*

Cooke, N. J., Gorman, J. C., Duran, J., Myers, C. W., & Andrews, D. (2013). Retention of team coordination skill. In W. Arthur, Jr., E. A. Day, W. Bennett, Jr., & A. Portrey (Eds.), *Individual and team skill decay: The science and implications for practice,* (pp. 344-363). New York: Taylor & Francis/Psychology Press.

10.3 Conference Proceedings

Demir, M., McNeese, N., Cooke, N., & Myers, C. (in press). The Synthetic Teammate as a Team Player in Command-and-Control Teams. *2016 Annual Meeting of Human Factors and Ergonomic Society.* Washington D.C. Human Factors and Ergonomics Society.

McNeese, N. & Cooke, N. (In Press). Team Cognition As A Mechanism For Developing Collaborative and Proactive Decision Support in Unmanned Aerial Systems. *18th International Conference on Human- Computer Interaction.* Toronto, CA.

Demir, M., McNeese, N. J., & Cooke, N. J. (2016). Team Communication Behaviors of the Human Automation Teaming. *Proceedings of the 2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support.* *Winner of best paper award.

Demir, M., McNeese, N., Cooke, N., Ball, J, Myers, C. (2015). Synthetic Teammate Communication and Coordination with Humans. Proceedings of the 59th Annual Conference of the Human Factors and Ergonomics Society, Santa Monica, CA: Human Factors and Ergonomics Society (pp. 951-955).

Demir, M., & McNeese, N. (2015). The Role of Recognition Primed Decision Making in Human Automation Teaming. *International Conference on Naturalistic Decision Making 2015.* McLean, VA.

Cooke, N. J., Bennett, W., Dougherty, J., Gawron, V., Neville, K., Rowe, L. & Shattuck, L. (2014). Panel: Remotely Piloted Aircraft Systems: A Human Systems Integration Perspective. *Proceedings of the 58th Annual Conference of the Human Factors and Ergonomics Society,* Santa Monica, CA: Human Factors and Ergonomics Society.

Demir, M. & Cooke, N. J., (2014). Human teaming changes driven by expectations of a synthetic teammate. *Proceedings of the 58th Annual Conference of the Human Factors and Ergonomics Society,* Santa Monica, CA: Human Factors and Ergonomics Society.

Cooke, N. J. & Myers, C. W. (2013). Issues relevant for synthetic teammate-human teammate interactions in operations of a synthetic unmanned aerial system.. In proceedings of the 17th International Symposium on Aviation Psychology. Dayton, OH: Wright State University.

10.4 Presentations

Cooke, N.J. & Myers, C. W. (2013). Issues relevant for synthetic teammate-human teammate interactions in operations of a synthetic unmanned aerial system. Paper presented at the

17th International Symposium on Aviation Psychology. May 6-9, Dayton, OH: Wright State University.

Cooke, N. J. (2013). Synthetic Task Environments: Challenges and Opportunities. Paper presented at the 17th International Symposium on Aviation Psychology. May 6-9, 2013, Dayton, OH: Wright State University

APPENDIX A

Effective Communication with Synthetic Teammate for DEMPC

A good way to achieve effective communication is to communicate using messages that are unambiguous and concise, without being cryptic. As the DEMPC, you are responsible for communicating information about the sequence of waypoints that are to be visited, to the AVO, during the course of a 40 minute mission. For each waypoint, you should communicate the name and type of the waypoint. You should also communicate any airspeed or altitude restrictions. Finally, you should communicate the effective radius. Here is a sample text message that communicates all this information:

DEMPC to AVO: The first waypoint is SA. It is an entry. The airspeed restriction is from 50 to 200. There is no altitude restriction. The effective radius is 5.

DEMPC to AVO: The next waypoint is TKE. It is a target. The airspeed restriction is from 50 to 200. There is no altitude restriction. The effective radius is 5.

The first sentence identifies and names the waypoint. The second sentence specifies the type of the waypoint. The third sentence specifies the airspeed restriction. The fourth sentence notes that there is no altitude restriction. The last sentence conveys the effective radius. All this information is needed by the AVO to perform his or her piloting task.

For the purposes of this experiment, you should not assume that the AVO and PLO are native speakers of English. There may be limitations in their understanding of English. For this reason, you should avoid highly cryptic and esoteric language. For example, the above information could have been provided as:

DEMPC to AVO: H-area=target. A=50-200. No alt. restr. R=5.

Although this message conveys most of the same information, due to its cryptic nature, it is difficult to understand. What do the 3 uses of '=' mean in this message? The abbreviation 'A' for airspeed is ambiguous with 'A' for altitude. 50-200 is not identified as a restriction. The abbreviation 'R' for radius is ambiguous with restriction. The abbreviations 'alt.' and 'restr.' might also be confused.

Besides avoiding cryptic and ambiguous language, it is best to convey all the information in a single message. If this is not done, then messages from the PLO to the AVO may interrupt your message and cause confusion. For example, consider the following sequence of messages sent to the AVO:

DEMPC to AVO: The next waypoint is SEL. It is an exit.
PLO to AVO: Raise altitude above 3000.
DEMPC to AVO: There are no restrictions.

If the PLO's message is about the current waypoint H-area, and not the next waypoint SEL, then the AVO may be confused and assume that the altitude restriction applies to SEL. If so, the AVO will be further confused by the next message stating there are no restrictions.

In addition to avoiding cryptic abbreviations, it is also a good idea to avoid misspellings which might confuse the AVO or PLO. For example, consider

DEMPC to AVO: Go to haere next!

The likelihood of the AVO recognizing that 'haere' should be 'H-area' is not good. This is especially true since this misspelling is close to the familiar word 'here' which will interfere with recognition of the less familiar waypoint name 'H-area'.

Effective Communication with Synthetic Teammate for PLO

A good way to achieve effective communication is to communicate using messages that are unambiguous and concise, without being cryptic. As the PLO, you are responsible for communicating information about the photo restrictions for each target waypoint that is to be visited, to the AVO, during the course of a 40 minute mission. For each target waypoint, you should communicate the name of the waypoint and the photo restriction. Here are a couple of sample text messages that communicate this information:

PLO to AVO: Raise altitude above 3000 for H-area.
PLO to AVO: Lower altitude below 3000 for F-area.

You are also responsible for notifying the AVO when a photo has been taken. Here's a sample:

PLO to AVO: Got the photo. Let's go.

You may also want to encourage the AVO to go faster so more photos can be taken:

PLO to AVO: Go faster above 300 for H-Area.

In some circumstances (e.g. when waypoints are close together), you may need more time to take a photo, in which case you may want to communicate the opposite:

PLO to AVO: Go slower below 200 for F-Area.

If you fail to get a photo of a target waypoint, you may need to request that the AVO return to the waypoint:

PLO to AVO: Go back to H-area.

For the purposes of this experiment, you should not assume that the AVO and DEMPC are native speakers of English. There may be limitations in their understanding of English. For this reason, you should avoid highly cryptic and esoteric language. For example, the photo restriction could have been provided as:

PLO to AVO: A>3000

Although this message conveys information about a photo restriction, due to its cryptic nature, it is difficult to understand. The abbreviation 'A' for altitude is ambiguous with 'A' for airspeed. > 3000 is not identified as a restriction nor is 3000 identified as being measured in feet. The waypoint that this restriction applies to is not mentioned.

In addition to avoiding cryptic abbreviations, it is also a good idea to avoid misspellings which might confuse the AVO or PLO. For example, consider

PLO to AVO: Got photo. Go to haere!

The likelihood of the AVO recognizing that 'haere' should be 'H-area' is not good. This is especially true since this misspelling is close to the familiar word 'here' which will interfere with recognition of the less familiar waypoint name 'H-area'.